Overlapped Trajectory-Enhanced Visual Tracking

This is the appendix to the paper entitled 'Overlapped Trajectory-Enhanced Visual Tracking', submitted to IEEE Transactions on Circuits and Systems for Video Technology.

APPENDIX A
SUPPLEMENTARY ABLATION STUDY

The ablation study on all datasets are shown in Table A.1 and A.2. It can be observed that the phenomena of the experiment results are analogous to those in the main text. All proposed methods show their effectiveness or efficiency. Specifically, 'w MP' is still the only variant that can outperform OTETrack (The performances of OTETrack and 'w/o ANA' are very close.). However, the performance gains are minor in all occasions and they are not worth the high extra computation expenses when compared with OTETrack.

APPENDIX B
SUPPLEMENTARY PARAMETER SENSITIVITY

We supplement additional parameter sensitivity analysis in this section. The involved hyper-parameters are the matching score threshold (η), the trajectory length (h), the update interval (γ), the smoothing parameters (b_1, b_2) and the peak area size coefficient of LH window (β). The parameter sensitivity of them are evaluated based on the attributes of LaSOT. Thus, more insights about their parameter sensitivity in varying tracking conditions can be provided. It can be observed that:

1) The increasing of matching score threshold η yields a less frequent dynamic template update and a more frequent use of SES. Therefore, a more strict threshold to replace the dynamic template with the previous tracking result leads to less spurious temporal information and more reliable dynamic templates. It can be observed from Fig. B.1(a) that setting η large makes OTETrack performs better when handling partial tracking scenarios where the spatial template matching ability and reliability are more significant, e.g., the ones with viewpoint change and rotation. However, the preference to SES results in an insufficient use of historical trend, making the performance of OTETrack with larger η gets worse when handling partial tracking scenarios where the temporal trajectory prediction is helpful for locating the target, e.g., the ones with background clutter and partial occlusion. On contrary, setting η too small leads to an opposite phenomenon. Therefore, using a moderate η is quite important for OTETrack to achieve the best performance and setting η = 0.6 achieves the most balanced performance among all tracking scenarios. Since η cannot be selected intelligently, it is one of the drawbacks of this work and our future research direction, as claimed in the last section of the main text.
2) The results in Fig. B.1(b) illuminates the robustness of our proposed trajectory prediction method with ES and LH window. Obviously, the prolonged trajectory length enhances the model tracking performance, especially when handling the tracking scenarios with fast motion and partial occlusion. Moreover, the tracking performance does not degrade even setting h = 10, which is rarely used by other trackers. These phenomena demonstrate the success of non-parametric trajectory prediction for single-object visual tracking and the validity of our

TABLE A.1
ABLATION STUDY OF THE SPATIAL PARTS ON ALL DATASETS

Models	GOT-10k			TrackingNet			LaSOT			$\text{LaSOT}_{\text{ext}}$			UAV123		
	AO(%)	$\text{SR}_{0.5}$(%)	$\text{SR}_{0.75}$(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)
OTETrack_{256}	76.4	85.4	75.1	84.8	89.3	83.9	73.9	83.5	82.3	51.9	62.3	59.2	70.8	86.3	91.3
w/o OViT	75.1	84.8	72.3	83.7	88.4	82.8	71.2	81.8	78.6	50.9	61.9	57.6	68.9	84.7	89.5
OTETrack_{256}-h	75.8	84.9	73.1	84.0	88.6	83.2	72.5	82.1	79.9	51.2	62.1	58.7	70.2	85.7	90.7
w/o ANA	76.6	85.5	75.3	84.9	89.4	84.2	73.8	83.6	82.2	51.8	62.2	59.2	70.9	86.3	91.4
w MP	76.8	85.6	75.2	85.1	89.5	84.3	74.1	83.6	82.3	52.1	62.5	59.3	71.0	86.4	91.5

TABLE A.2
ABLATION STUDY OF THE TEMPORAL PARTS ON ALL DATASETS

Models	GOT-10k			TrackingNet			LaSOT			$\text{LaSOT}_{\text{ext}}$			UAV123		
	AO(%)	$\text{SR}_{0.5}$(%)	$\text{SR}_{0.75}$(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)	AUC(%)	P_{Norm}(%)	P(%)
OTETrack_{256}	76.4	85.4	75.1	84.8	89.3	83.9	73.9	83.5	82.3	51.9	62.3	59.2	70.8	86.3	91.3
w/o ES	75.8	84.8	72.9	83.9	88.5	82.9	71.0	81.5	78.3	50.1	60.5	56.2	69.2	84.7	89.1
w/o HES	76.2	85.2	73.9	84.2	88.6	83.4	73.2	83.0	80.9	51.0	62.0	57.9	70.4	86.0	90.8
w/o SES	75.9	85.0	73.4	84.0	88.5	83.2	71.4	81.6	78.5	50.5	60.8	56.7	70.2	85.3	90.6
w/o LH	76.1	85.1	74.9	84.3	88.7	83.4	72.1	81.4	79.8	51.0	61.9	58.0	69.4	84.8	89.4

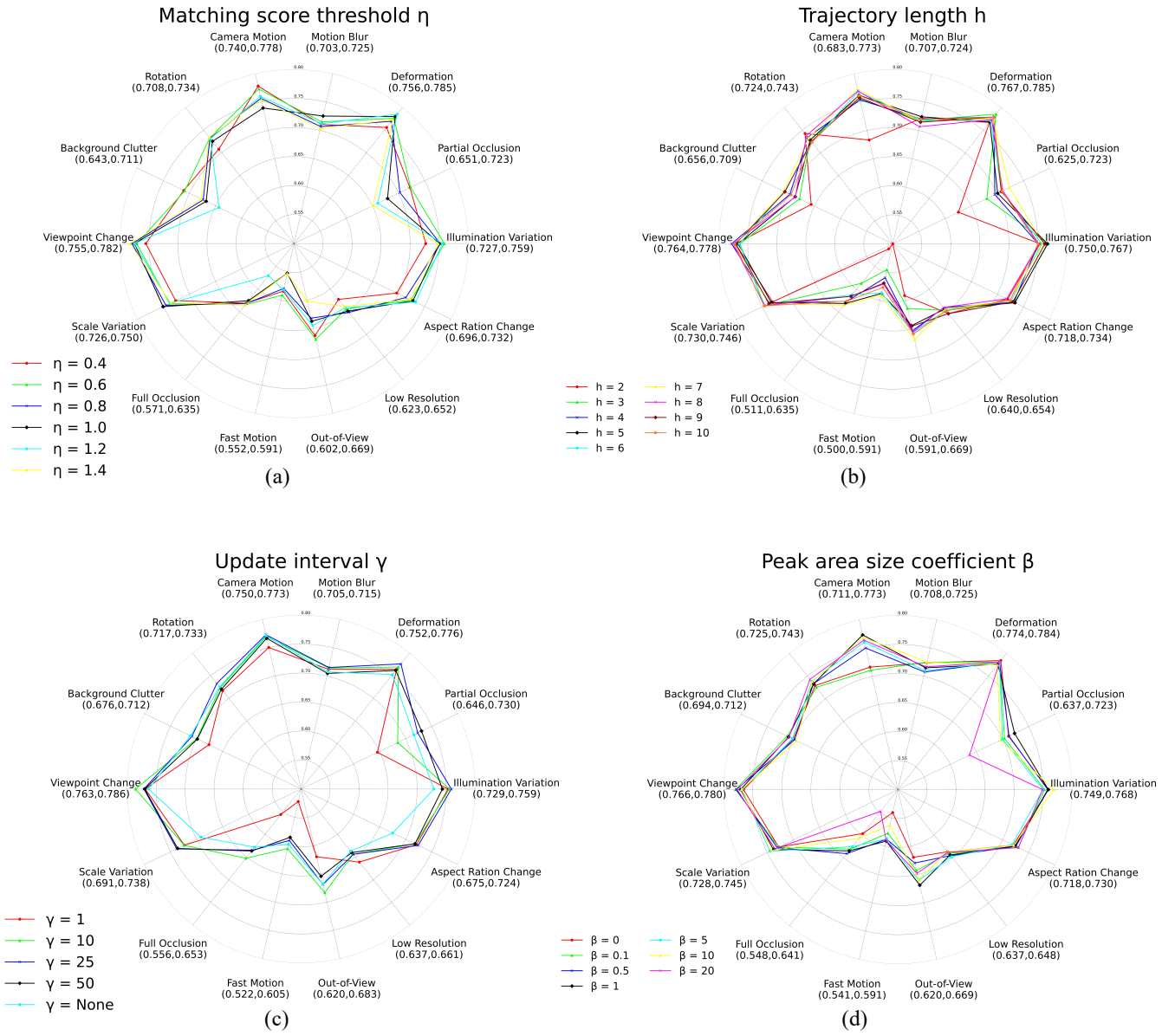

Fig. B.1. The parameter sensitivity analysis upon the attribute-based evaluation on LaSOT. (a) The parameter sensitivity of matching score threshold η. (b) The parameter sensitivity of trajectory length h. (c) The parameter sensitivity of update interval γ. (d) The parameter sensitivity of peak area size β.

proposed method.

3) The update interval is an important hyper-parameter for those trackers [1], [2] adopting the dynamic template update method. However, the performance of OTETrack is quite stable with different γ, except the extreme conditions where $\gamma = 1$ or the dynamic template is not applied ($\gamma = \text{None}$). These show that the template matching ability of OTETrack is qualified thanks to the strong Overlapped ViT backbone.
4) In Fig. B.1(d), the performance of OTETrack gets worse if setting β too large or too small. Since the small β makes the LH window approach the conventional Hanning window, the tracking performance of OTETrack with small β is prone to the motion. Therefore, OTETrack with small β performs worse than the ones with larger β when handling the tracking scenarios with fast motion or camera motion. On contrary, an extremely large β renders LH window useless and the temporal information not applied to OTETrack. Thus, OTETrack with larger β owns much worse performance than the ones with smaller beta under the occasions where the background clutter or occlusion occurs. However, it can also be observed that changing β does not make significant differences in most of the cases if its value is in an appropriate range, which demonstrates that using the standard deviation of trajectory to control the peak area size of LH window is feasible and robust.

Specifically, the parameter sensitivity of smoothing parameters is visualized to highlight the reason of applying ES method as the prediction method, i.e., its property of laying more

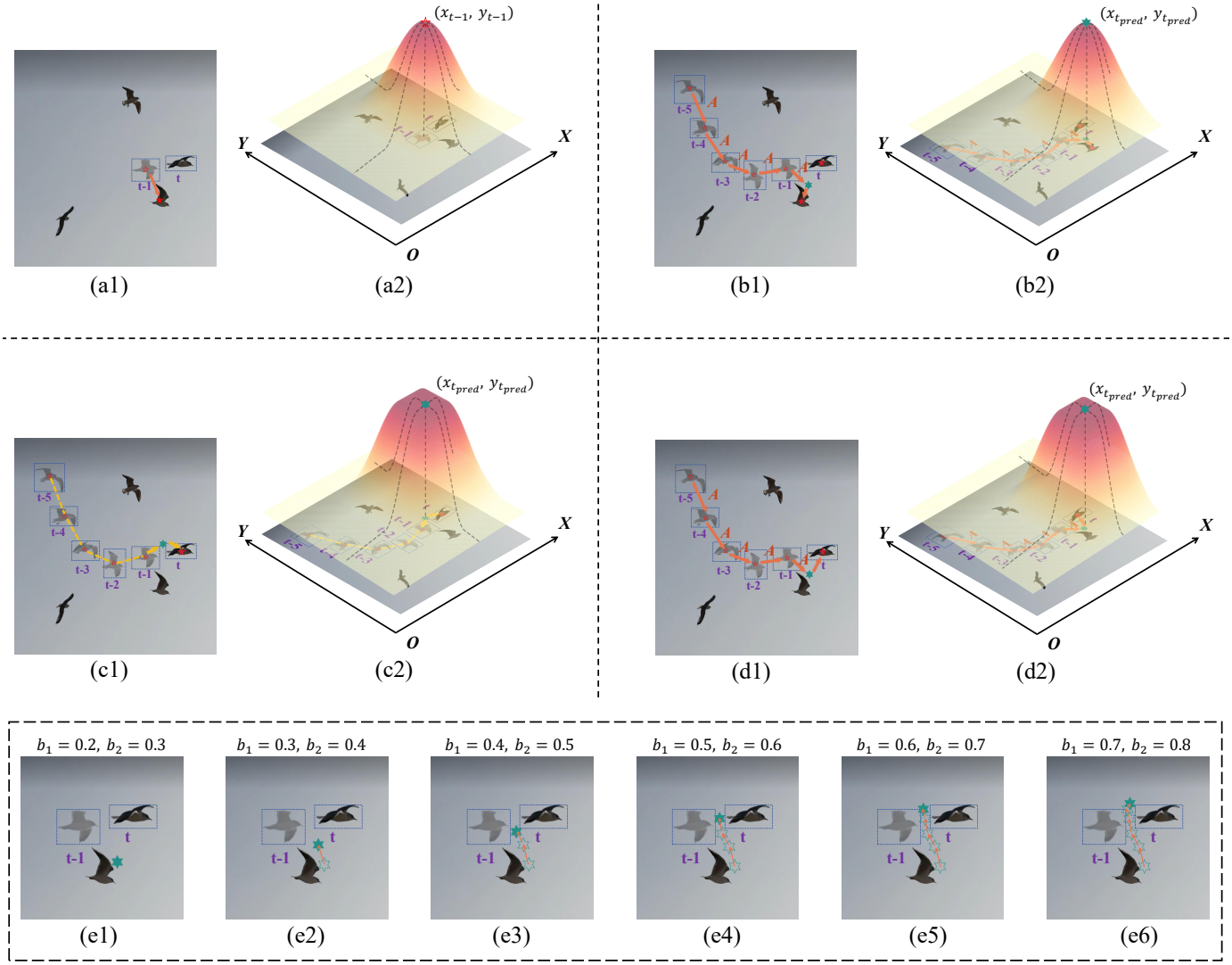

Fig. B.2. The visualization of the OTETrack variants with various trajectory prediction methods. The target of this tracking scenario is to trace a specific bird. The transparent birds represent the target locations of the historical frames (from $t-5$ to $t-1$). The red pentagrams denote the true center positions of the targets during the trajectory. The green hexagrams denote the prediction results of the current target positions (at t). A represents the parametric transformation between the target positions of adjacent frames. (a) OTETrack replacing LH window with conventional Hanning window. (b) The ARTrack with conventional Hanning window. (c) The original and complete OTETrack. (d) The ARTrack with LH window. Each of the above four figure groups contains two sub-figures. The first one is to highlight the trajectory and the second one is to highlight the influences of Hanning window or LH window. (e) The HES trajectory prediction results with varying b_1 and b_2.

emphasis on the later trajectory elements. We merely visualize the performance of HES and its hyper-parameters b_1, b_2 in that the basic concepts of HES and SES are similar. Moreover, the corner coordinate prediction is replaced with center prediction to present a more vivid visualization. As shown in Fig. B.2(a), the conventional Hanning window, which sets the position of the target in the previous frame (at $t-1$) as the window center, makes the tracker follow the wrong object as the location of the background bird is more close to the Hanning window center than that of the target bird. Therefore, it is significant to predict the current position (at t) of the target and use it as a more accurate Hanning window center. Unfortunately, though the tracker in Fig. B.2(b) possesses a module to predict the current position of the target, it is done by the recursive parametric prediction. Therefore, it fails to handle the occasion in Fig. B.2 where the target suddenly changes the direction (from $t-2$ to $t-1$) and still traces the wrong target due to the distribution shift of the trajectory. In contrast, our ES-based prediction method helps the tracker instantly capture the abrupt motion change of the target and makes the latest position change dominate the trajectory prediction result. Therefore, the tracker using our proposed method successfully follow the target in Fig. B.2(c). Moreover, the usage of LH window further improves the probability of a successful tracking. As shown in Fig. B.2(d), the tracker, which applies recursive parametric prediction to the trajectory like the one in Fig. B.2(b) but uses the LH window approach, also successfully follows the correct target. Thanks to the peak area of LH window, which is quite large due to the fast motion of the target, the positions of the background bird and the target bird

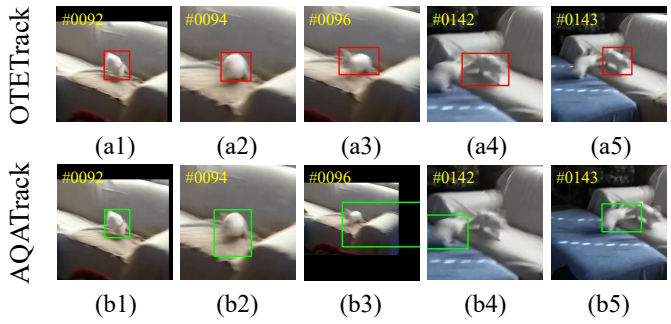

Fig. C.1. Visualization of OTETrack and AQATrack on rabbit-10 of LaSOT, where motion blur problem occurs. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of AQATrack.

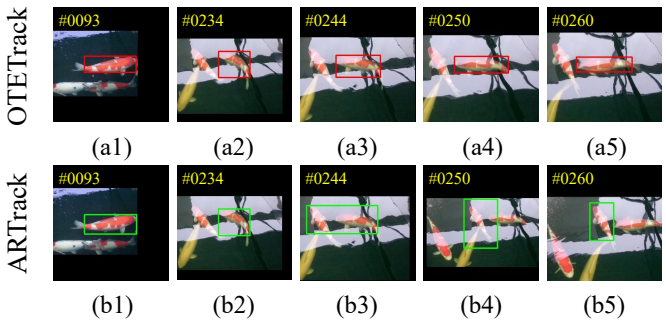

Fig. C.3. Visualization of OTETrack and ARTrack on goldfish-8 of LaSOT, where illumination variation problem occurs. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of ARTrack.

possess the identical value in LH window. Thus, the tracker may follow the target if its spatial template matching ability is qualified.

Besides, we also present the trajectory prediction results of HES with different b_1 and b_2 in Fig. B.2(e). With the prolonged b_1 and b_2, which make HES more focuses on the later trajectory elements to make the prediction, the predicted target position in the current frame is more close to the ground truth. Therefore, appropriately large ES hyper-parameters help the tracker better capture the distribution shift of the trajectory. Moreover, when setting $b_1 \geq 0.6$ and $b_2 \geq 0.7$, the prediction results slightly change. Therefore, we set $b_1 = 0.7$ and $b_2 = 0.8$ in the experiment of the main text. Undoubtfully, flexibly setting these two parameters according to the real-world application scenarios would further enhance the performance of OTETrack, indicating that OTETrack can perform even better in real-world practice.

APPENDIX C
SUPPLEMENTARY CASE STUDY

C.1 *Showcases of Public Benchmarks*

Since the case studies in the main text merely include the tracking occasions involving background clutter and out-of-view problems, we provide more case studies, which are involved with other types of tracking scenarios, to more comprehensively highlight the tracking ability and robustness of OTETrack:

1) Fig. C.1 illustrates the rabbit-10 sequence of LaSOT where motion blur problem occurs. As the motion blur makes the target location change fast and the target appearance ambiguous, such tracking scenario both tests the model spatial template matching capability and the temporal information application capability. We use AQATrack [3], which is second only to OTETrack on motion blur according to Fig. 6 of the main text, for comparison. As shown in Fig. C.1(b), AQATrack fails to appropriately apply the temporal information so that it starts to lose its target at the 96th frame where the target abruptly moves. Then, AQATrack eternally loses the target since its spatial template matching capability is also not sufficient to assist it in tracing back the ambiguous target. However, OTETrack succeeds in tracing the target during the time span from Fig. C.1(a1)-(a5), demonstrating its superb tracking capability.
2) Another tracking scenario involved with viewpoint change is sketched in Fig. C.2. These frames belong to the train-1 sequence of LaSOT. Similarly, EVPTrack [4] is used as the competitor due to its second best performance in Fig. 6 of the main text. The tracking scenarios with respect to viewpoint change problem depict the target and its background from varying viewpoints so that they solely examine the model spatial templating matching capability. Therefore, thanks to the strong and efficient feature extraction capability of Overlapped ViT,

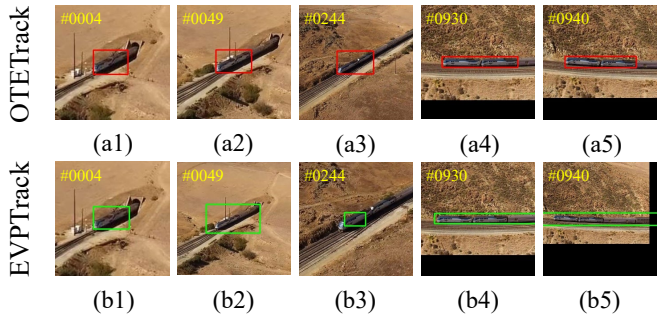

Fig. C.2. Visualization of OTETrack and EVPTrack on train-1 of LaSOT, where viewpoint change problem occurs. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of EVPTrack.

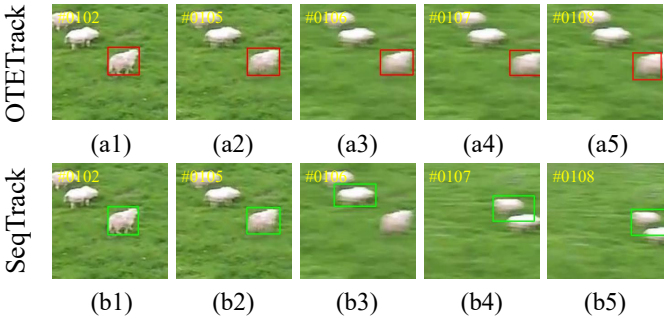

Fig. C.4. Visualization of OTETrack and SeqTrack on sheep-9 of LaSOT, where camera motion problem occurs. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of SeqTrack.

OTETrack is more competent in handling viewpoint change problem compared with EVPTrack in Fig. C.2.

3) The illumination variation can be treated as adding annoying and non-negligible noises to the search image, thus the target features are severely polluted or even covered. Obviously, The illumination variation problem is even more challenging than viewpoint change for the spatial template matching capability of trackers. Analogously, OTETrack outperforms other trackers in this occasion, including the second best ARTrack [5], which loses the target in the 244th frame, in Fig. C.3.
4) Fig. C.4 illustrates the sheep-9 sequence of LaSOT where camera motion occurs. Camera motion is a common problem in reality because the camera carrier can never keep steady. In such cases, the target appearance is more vague than normal and its location is casual. Therefore, the spatial template matching capabilities of trackers are tested. Moreover, the trackers with the conventional Hanning windows have difficulty in handling such scenarios since the conventional Hanning window is prone to the casual motion. For instance, SeqTrack [1] fails to trace the correct target when the camera motion happens in the 106th frame, as shown in Fig. C.4(b3). In contrast, the camera motion does not affect OTETrack even a little in Fig. C.4(a), which demonstrates the strong template matching capability of OTETrack and the effectiveness of its trajectory prediction method.

Besides, we also present some showcases involving the low tracking qualities and the tracking failures of OTETrack as follows:

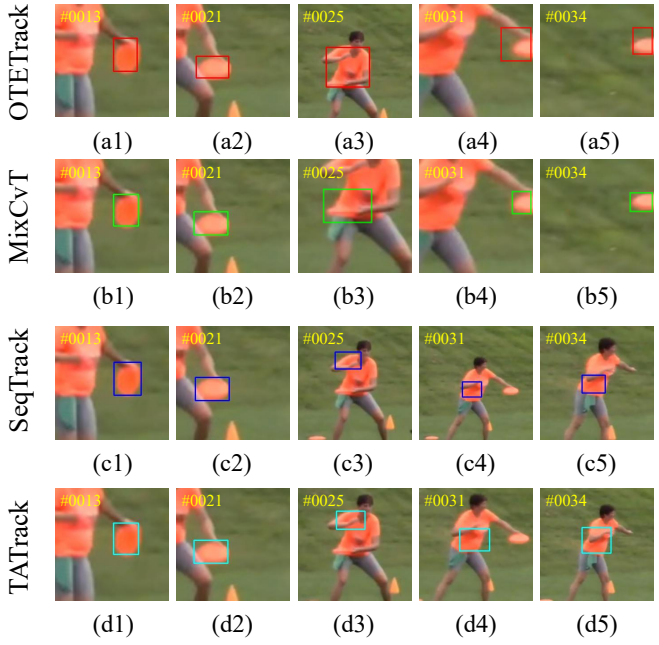

Fig. C.5. Visualization of OTETrack, MixCvT, SeqTrack and TATrack on frisbee-7 of LaSOT$_{\text{ext}}$, where fast motion, out-of-view and low resolution problems occur. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of MixCvT. (c) The results of SeqTrack. (d) The results of TATrack.

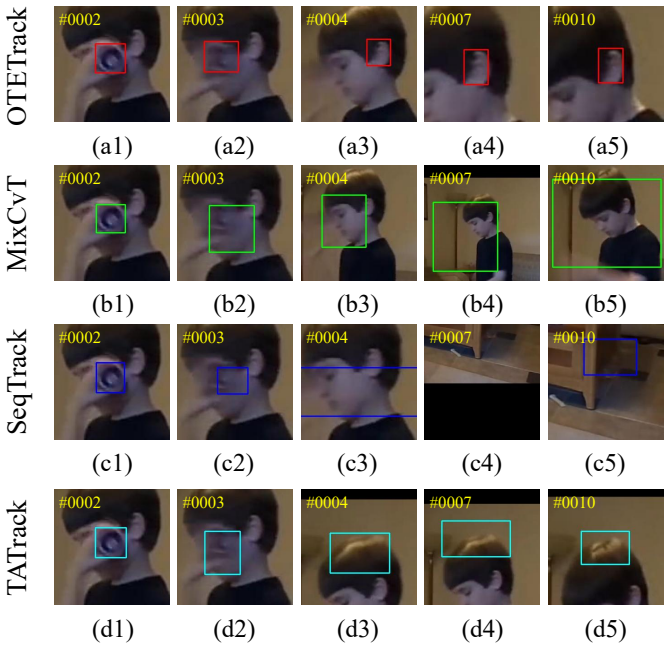

Fig. C.6. Visualization of OTETrack, MixCvT, SeqTrack and TATrack on yoyo-7 of LaSOT$_{\text{ext}}$, where extreme fast motion, out-of-view and low resolution problems occur. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of MixCvT. (c) The results of SeqTrack. (d) The results of TATrack.

1) The target in the frisbee-7 sequence of LaSOT$_{\text{ext}}$ is a fast-moving small frisbee, which means that the frisbee-7 sequence possesses the attributes of fast motion and out-of-view. Specifically, in the tracking scenarios of Fig. C.5(a3) and (b3), trackers shall additionally face the low resolution problem since the color of the background is analogous to the target. Benefitting from bigger search image size (320 vs. 256), MixCvT [2] handles such tracking scenarios better than OTETrack. However, OTETrack stills does not lose the target in Fig. C.5, making it perform better than the other trackers with the search image sizes of 256 [1], [6] in Fig. C.5.
2) Similarly, the tracking scenarios in Fig. C.6 also possess the attributes of fast motion, out-of-view and low resolution, but are more extreme. The target yoyo moves so fast that it completely leaves the view in the third frame and not comes back. None of trackers could follow the target in such situations due to the limited view size. This is the drawback of cropping the search image, which is commonly adopted and is supposed to improve the efficiency. In practice, the users of trackers may increase the cropped search image size to handle this problem.

C.2 *Showcases of Real-world Application*

To show that OTETrack is literally practical in real-time applications, we use an Unmanned Aerial Vehicle (UAV), whose product model is DJI Mini 2, with a NVIDIA Jetson AGX Xavier[1] onboard computer to monitor a special person

[1]https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-series/.

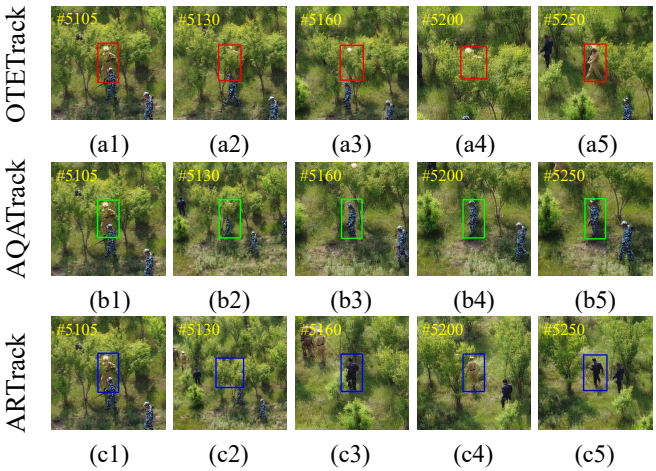

Fig. C.7. Visualization of OTETrack, AQATrack and ARTrack on a real-world application to trace a special person, where occlusion and background clutter problems occur. The frame indexes are in the frames. (a) The results of OTETrack. (b) The results of AQATrack. (c) The results of ARTrack.

in a group of people. These people keep moving in single file and they are dressed similarly. The circumstance is full of trees and other obstacles, which may occlude the people from time to time. The above conditions show that this tracking scenario at least faces the problem of background clutter and occlusion. However, as shown in Fig. C.7, OTETrack still handles this tracking scenario with better accuracy than other baselines. Specifically, both AQATrack and ARTrack lose their targets when the occlusion occurs in the 5130th frame. The AQATrack traces a wrong person that is near to the target. ARTrack traces a wrong person that is several steps forward due to its recursive parametric prediction method. These phenomena demonstrate the superb and robust tracking ability of OTETrack.

REFERENCES

[1] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu, "Seqtrack: Sequence to sequence learning for visual object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 572–14 581.

[2] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Mixformer: End-to-end tracking with iterative mixed attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 608–13 618.

[3] J. Xie, B. Zhong, Z. Mo, S. Zhang, L. Shi, S. Song, and R. Ji, "Autoregressive queries for adaptive tracking with spatio-temporaltransformers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[4] L. Shi, B. Zhong, Q. Liang, N. Li, S. Zhang, and X. Li, "Explicit visual prompts for visual object tracking," in *AAAI Conference on Artificial Intelligence*, 2024.

[5] X. Wei, Y. Bai, Y. Zheng, D. Shi, and Y. Gong, "Autoregressive visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9697–9706.

[6] K. He, C. Zhang, S. Xie, Z. Li, and Z. Wang, "Target-aware tracking with long-term context attention," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 773–780, 2023.